UNIVERSIDADE DE BRASÍLIA
FACULDADE DE ADMINISTRAÇÃO, CONTABILIDADE, ECONOMIA E GESTÃO DE POLÍTICAS PÚBLICAS
DEPARTAMENTO DE ECONOMIA

FELIPE CARNEIRO DE FIGUEREDO

# Ideology as a Belief System:
# A Computational Linguistic Approach to Brazilian Senate

Orientador: Prof. Dr. Bernardo Mueller

BRASÍLIA
2019

Prof. Dr. Márcia Abrahão Moura
Reitora da Universidade de Brasília

Prof. Dr. Eduardo Tadeu Vieira
Diretor da Faculdade de Administração, Contabilidade, Economia e Gestão de Políticas Públicas

Prof. Dr. Milene Takasago
Chefe do Departamento de Economia

FELIPE CARNEIRO DE FIGUEREDO

# Ideology as a Belief System:
# A Computational Linguistic Approach to Brazilian Senate

Dissertação apresentada como requisito parcial para a obtenção do título de Mestre em Economia pela Universidade de Brasília.

Orientador: Prof. Dr. Bernardo Mueller

BRASÍLIA
2019

Figueredo, Felipe Carneiro de
Ideology as a Belief System: A Computational Linguistic Approach to Brazilian Senate / Felipe Carneiro de Figueredo – Brasília, 2019.
39f.: il.; 30 cm

Dissertação apresentada como requisito parcial para a obtenção do título de Mestre em Economia pela Universidade de Brasília. – Universidade de Brasília
Orientador: Mueller, Bernardo

1. Ideologia. 2. Linguagem. 3. Aprendizado de Máquinas. 4. Ciência Política.

FELIPE CARNEIRO DE FIGUEREDO

**Ideology as a Belief System:**
**A Computational Linguistic Approach to Brazilian Senate**

Dissertação apresentada como requisito parcial para a obtenção do título de Mestre em Economia pela Universidade de Brasília.

Área de Concentração: Economia

**Data de Aprovação:**
22/03/2019

**Banca Examinadora:**

---

**Prof. Dr. Bernardo Mueller**
Orientador

---

**Prof. Dr. Daniel Cajueiro**
Avaliador Interno

---

**Prof. Dr. Lucio Rennó**
Avaliador Externo

Dedico esta dissertação para minha **família**,
aquela que sempre me aconselhou o caminho do
conhecimento.

## ACKNOWLEDGMENTS

Agradeço primeiramente à minha família por todo o apoio e solidariedade durante minha formação. Aos meus pais, irmã, tios, primos e avós, minha mais profunda gratidão.

Agradeço também a todo corpo docente do Programa de Pós-Graduação em Economia por toda excelência e profissionalismo prestado. Em especial, agradeço aos professores, Bernardo Mueller e Daniel Cajueiro, pela orientação e por me ensinarem o verdadeiro valor da ciência.

Agradeço também a todos meus colegas do Mestrado, os quais foram essenciais para meu sucesso. Em especial aos meus amigos Lucas Naves, Pedro Campelo e Matheus Borghi.

Por fim, agradeço ao CNPq pelo financimento da minha bolsa.

**ABSTRACT**

In this paper, we intend to analyze the patterns subsumed in the ideology. Ideology can be seen as a political belief system, intuitively expressing a view of which issues positions go together. We use a computational technique that involves classification regarding these constraints. The object of this analysis is the Federal Senate of Brazil from the 50th to the 55th legislatures, in which we analyze political speech to extract the dictionaries that best translate the content of each ideology. We also provide a dimensionality reduction technique in order to better identify how these data are arranged in a high-dimensional space. Finally, we further investigate the political dynamic among the legislatures comparing our results with current political science literature.

**Keywords**: ideology. language. machine learning. political science.

## RESUMO

Neste trabalho, pretendemos analisar os padrões subordinados na ideologia. A ideologia pode ser vista como um sistema de crença política, expressando intuitivamente uma visão de quais posições sobre determinadas questões caminham juntas. Usamos uma técnica computacional que envolve a classificação em relação a essas restrições. O objeto dessa análise é o Senado Federal do Brasil, da 50ª à 55ª legislatura, em que analisamos o discurso político para extrair os dicionários que melhor traduzem o conteúdo de cada ideologia. Também fornecemos uma técnica de redução de dimensionalidade para melhor identificar como esses dados são organizados em um espaço de alta dimensão. Finalmente, investigamos ainda mais a dinâmica política entre as legislaturas comparando nossos resultados com a literatura atual de ciência política.



**Palavras-chave**: ideologia. linguagem. aprendizado de máquinas. ciência política.

## LIST OF FIGURES

## LIST OF TABLES

## LIST OF ABBREVIATIONS AND ACRONYMS

| | |
|---|---|
| CPI | Comissão Parlamentar de Inquérito |
| ESEB | Estudo Eleitoral Brasileiro |
| KL | Kullback-Leibler |
| NLP | Natural Language Processing |
| PCA | Principal Component Analysis |
| PCdoB | Partido Comunista do Brasil |
| PEC | Proposta de Emenda à Constituição |
| PP | Partido Progressista |
| PSDB | Partido da Social Democracia Brasileira |
| PT | Partido dos Trabalhadores |
| PTB | Partido Trabalhista Brasileiro |
| SVD | Singular Value Decomposition |
| tf-idf | Term Frequency Inverse Document Frequency |
| t-SNE | t-Distributed Stochastic Neighbor Embedding |

## SUMMARY

# INTRODUCTION

In contemporary politics, we perceive the political debate as an actual battlefield. Politicians from all across the political spectrum engage in several battles about topics such as abortion, immigration, gun control, gay marriage, and taxation. Most of these disputes reveals an intrinsic challenge underlying our most general and stable political system, democracy. The discussion, then, turns towards its disruption. Is democracy failing? In a recent book, Levitsky & Ziblatt (2018) argue that politicians now treat their rivals as enemies, intimidate the free press, and threaten to reject the results of elections. Still according to them, politicians now try to weaken the institutional buffers of our democracy, including the courts, intelligence services, and ethics offices. Surely, one indicative sign of this current crisis is recent political polarization – many times revealed as ideological disputes. Hare & Poole (2014) and Gentzkow et al. (2019) analyzing the United States' political context, find a drastic movement towards polarization [1]. Meanwhile, Harari (2018) suggests that the most revealing inflection point was the year 2016 – marked by the Brexit vote in Britain and the rise of Donald Trump in the United States – "signifying the moment when this tidal wave of disillusionment reached the core liberal states of western Europe and North America". In the Brazilian context, former president Dilma's impeachment process followed by the presidential victory of far-right politician, Jair Bolsonaro, reveals an extremely bellicose scenario. Understanding this contemporary dynamic has become an intricate and challenging job.

In order to assess the nature and extension of contemporary polarization, we need some method for the measurement of ideology. However, the measurement of ideology is one of the most difficult tasks in political science, mostly because it is not directly observable. In this effort, many scholars such as Poole & Rosenthal (1985) and Power & Zucco (2009) employ different strategies, ranging from survey responses to statistical estimates based on voting records. Their results rely upon a low-dimensional approach, so that every political thought can be projected into a spatial model. Thus, such constructs measure ideology as a discrete concept and actually are the main building block of political behaviour representations. This set of ideas, albeit, are not restricted across limited dimensions. Besides, there is certain interest in understanding the content of these positions.

We see that models which best represent idiosyncrasies between political parties are those based on multi-dimensional frameworks (Diermeier et al., 2011.) Because they can take in account a large number of features, instead of either economic or redistribution issues separately, they display a richer representation of political positions. Moreover, we know that one great source of high dimensionality is language. Through political speech we can analyze the patterns subsumed in the ideology labels. For instance, manually classifying phrases into sub-

[1] Although, these two works diverge about the extent of this phenomenon. Hare & Poole (2004) analyzing a standard measure of ideological polarization based on roll-call votes, find that political polarization in the US is currently as high it was in the late nineteenth and early twentieth century, and its current upward trend began around 1950 rather than 1990 – as found by Gentzkow et al. (2019.)

stantive topics shows that the increase in partisanship is due more to changes in the language used to discuss a given topic (e.g., “estate tax” vs. “death tax”) than to changes in the topics parties emphasize (Gentzkow et al., 2019.) These language-based measure also suggest that speech and roll-call votes respond to different incentives and constraints [2].

This dissertation is based on a novel approach that seeks to analyze and provide a general outlook about the content of ideologies. This method treats language as the unity of analysis, applying an automatic classification algorithm to represent dictionaries that best explain each political position. It provides, for instance, a model to study legislative behavior, intra-party politics, and polarization (Lauderdale and Herzog, 2016.)

Underlying this approach is the hypothesis that ideologies give coherence to a person’s opinions and attitudes, so that once we have properly identified a person’s ideology, we may be able to predict his or her opinions on new or modified issues. In contemporary politics the knowledge that a politician opposes the raising of corporate taxes and in the minimal wage makes him most likelly also favorable to a balanced budget, against affirmative action, etc. In other words, people hold their political views – even those that are logically organized – with passion (Poole, 2003.) [3]

Finally, our contribution is to help uncover the link between the content of each political position and its underlying ideology. In this effort, we depart from a predefined party classification[4] in that we build a political dictionary that evolves across a timeline. This framework is useful to analyze the stability and transition of beliefs, in an approach similar to Alston et al. (2013) who analyze the change in the social contracts in Brazil – although we rely upon a machine learning technique. Thus, using political speech data from the 50th legislature to the 55th legislature of the Federal Senate of Brazil (1995 - 2018), we analyze its dynamic so as to identify what drives the changes in polarization. Providing a novel approach to understanding Brazilian politics, it contributes to the current literature by enabling the investigation of several important hypotheses [5].

The outline of the dissertation is as follows. In Chapter 1, we present our model – describing in-detail how our data was prepared and what methods were employed. In Chapter 2, we present the general results of our classification method and we also analyze their extent. In doing so, we reveal a certain lack of variability with some punctuated discontinuities – translated as political drifts. Also, we make links between our results and the present political science literature. Lastly, in Chapter 3, we present a brief conclusion concerning our findings.

---

[2] Roll-call votes may be shaped by strategic considerations related to the passage of legislation, and may therefore not reflect legislators’ sincere policy preferences. Speech may reflect party differences in values, goals, or persuasive tactics that are distinct from positions on specific pieces of legislation (Gentzkow et al., 2019.)

[3] In short, as argued by Converse (1964), ideology is a belief system – namely, that issues are interrelated or bundled and that ideology is fundamentally the knowledge of "what-goes-with-what". He pointed that the sources of constraints on idea-elements are much less logical in the classical sense than they are psychologically — and less psychologically than in social contexts.

[4] Power and Zucco, 2009.

[5] In this work, we investigate, for example, the *direita evergonhada* (literally, the "ashamed right") hypothesis, and the historical ideological proximity between two of the most powerful parties in Brazil, PT and PSDB (Power and Zucco, 2009.)

# 1 THE MODEL

A wide variety of questions may be addressed with textual analysis. For this, we have manual and automated methods of analysis [1]. The advances of computational techniques enable recently the capacity of textual analysis from manual to automated methods. Usually, automated methods are more objective and enable researchers to analyze more texts in less time than traditional manual methods. Although, the complexity of language implies that automated content analysis methods will never replace careful and close reading of texts. Rather, these methods are best thought of as amplifying and augmenting careful reading and thoughtful analysis (Grimmer and Stewart, 2013.).

We use in our analysis standard automated techniques (i.e., Natural Language Processing, or NLP) in order to further understand each ideology – revealing and analyzing their content. In the literature, Diermeier et al. (2011) employed a similar approach [2]. Our model, then, relies upon a novel approach created by an intersection between machine learning and political science.

## 1.1 Literature Review

The first modern statistical analysis of text data, Mosteller & Wallace (1963), used text analysis to infer the authorship of the disputed *Federalist Papers* that had alternatively been attributed to either Alexander Hamilton or James Madison.

Text as data are widely employed as input for economic research. Scott & Varian (2015), for instance, attempt to "nowcast" macroeconomic variables using data on the frequency of Google search terms: search term counts are aggregated by week and by geographic location, then converted to location specific frequency indices. In a recent work, Kelly et al. (2019) use textual analysis of high-dimensional data from patent documents to create new indicators of technological innovation. According to them, these indices capture the evolution of technological waves over a long time span and are strong predictors of productivity at the aggregate, sectoral, and firm level.

In finance, others applications employing sentiment analysis through media analysis are a powerful method to forecast stock market activity (Tetlock, 2007; Boolen et al., 2011.) The text as data method also extends to cognitive science (Iter et al., 2018; Sagi and Diermeier, 2015; Sagi and Dehghani, 2013.) Iter et al. (2018), for instance, propose a novel computational model for referencing incoherence based on ambiguous pronoun usage and show that it is a highly predictive feature on schizophrenia.

In the field of political science, there are several attempts to analyze data from text (Laver et al., 2003; Yu et al., 2008; Diermeier et al., 2011; Jensen et al., 2012; Lauderdale et al., 2016;

[1] Graaf & van der Vossen (2013) make a comparative between them.

[2] Although they analyzed the United States Senate's speeches and they also applied a different classification algorithm, Support Vector Machines (SVM) – even though they employed our same classification setup as a baseline model.

Yan et al., 2018; Barron et al., 2018; Gentzkow et al., 2019.) For example, Gentzkow et al. (2019) analyze United States Congressional Record's data text from the 43rd Congress to the 114th Congress. They find that the partisanship of language has exploded in recent decades. While they cannot definitively determine why the partisanship of language increased when it did, the evidence points to innovation in political persuasion as a proximate cause [3]. While Barron et al. (2018) build a model to study novelty, transiency and ressonance on speech data from French Revolution.

## 1.2 Data Set

The data set of this analysis are the speech records from the Federal Senate of Brazil comprising its 50th legislature until its 55th legislature – ranging from 1995 to 2018 [4]. This interval represents an unique period in Brazil's recent democracy, because important debates – that drove important decisions – were then made. We can point to discussions about land reform, fiscal reform, pension reform, and one impeachment process [5].

We make use of all senators' political speech within this interval [6]. In doing so, we divide the entire data set regarding their legislatures. We therefore have one speech set (i.e., one document set, or *corpus*) for each legislature.

We treat each political speech as a document $d_i$ , such that $d_i \in \mathbb{R}^N$ $(i = 1, \ldots, M)$ – $N$ is the total number of words (or, dimensions.) and $M$ is the total number of speeches. For each document $d_i$, we have an unique category $c_i$, such that $c_i \in \{-1, 0, 1.\}$ Notice that we set only three categories: left (i.e, $c_{left} = -1$), center ($c_{center} = 0$), and right ($c_{right} = 1.$)

### *1.2.1 Speech Categorization*

The categories of the documents $c_i$ should be known in advance, typically by inspecting them individually and hand-labelling them (Bishop, 2006.) We follow the categorization employed by Power & Zucco (2009.) Their classification is widely used in the literature to study legislative behaviour, because they produced a very rich and in-depth analysis ranging from roll-call voting data to survey responses. They classify the main Brazilian political parties regarding their political position as left, center or right. For instance, the classification of three of the most important parties are: PT (Workers' Party, or *Partido dos Trabalhadores*) is located on the left (i.e, $c_{PT} = -1$), while PSDB (Brazilian Social Democracy Party, or *Partido da Social*

[3] The 1994 inflection point found in their data coincides with the Republican takeover of Congress led by Newt Gingrich, under a platform called the *Contract with America* (Gentzkow et al, 2019.) This election is widely considered a watershed moment in political marketing, with consultants such as Frank Luntz applying novel techniques to identify effective language and disseminate it to candidates (Lakoff, 2004.)

[4] We downloaded all content from the Brazilian Senate website.

[5] To analyze the senators speeches also provides us a real measure of how their political parties guidelines are build and evolve, because many of them compose the elite members from their respective parties.

[6] Our data set comprises approximately 80,000 speeches, including speeches and pronunciations. To have a point, in the Brazilian Senate each legislature has 81 senators (three for each of all 26 states plus the federal district.) Each senator holds a mandate for two legislatures.

*Democracia Brasileira*) is located on the center ($c_{PSDB} = 0$), and PP (Progressive Party, or *Partido Progressista*) is located on the right ($c_{PP} = 1$.)

### 1.3 Feature Selection

We can represent the documents $d_i$ as counting vectors [7]. Through that, we ignore the order of words such that $d_i$ is a vector whose length is equal to the $N$ number of words in the vocabulary and whose elements $w_{ij}$ are the number of times word $j$ occurs in document $d_i$.

We remove stop words as manner of reducing the complexity of our data (Jurafsky and Martin, 2018) [8]. Stop words are words that do not bear solid meaning (mostly function words, such as the, a, and of .)

Then, suppose that the text of document $d_i$ is [9]:

> E razões eleitorais não nos comovem. Vamos enfrentar isso. E acredito que é um risco que deve correr qualquer parlamentar, qualquer homem público. Contrariar aqui ou agradar acolá, isso é da vida pública.

After removing stop words, and removing punctuation, we might be left with "razões eleitorais comovem vamos enfrentar acredito risco deve correr qualquer parlamentar qualquer homem público contrariar agradar acolá vida pública". The bag-of-words representation would then have $w_{ij} = 2$ for $j \in$ {pública, qualquer}, $w_{ij} = 1$ for $j \in$ {razões; eleitorais; comovem; vamos; enfrentar; acredito; risco; deve; correr; parlamentar; homem, contrariar, agradar, acolá, vida}, and $w_{ij} = 0$ for all other words in the vocabulary.

In order to further reduce the dimensionality of our data, we set a minimum term frequency of 50, assuming that words with frequencies below that threshold have low coverage and thus are not useful for classification [10]. And, we also set a maximum term frequency of 95%, assuming that words with such high frequencies are not relevant for classification, because mostly of them do not bring meaningful elements in a language.

---

[7] The simplest and most common way to represent a document is through counting vectors (or, bag-of-words.) There are different ways to represent text as data: for instance, by providing a boolean representation, which is an algebraic expression consisting of binary values (1 or 0.) It assign a value of 1 for a simple occurrence of feature $j$ in document $i$, and when absent it assign the value of zero.

[8] Stemming is another manner of reducing the dimensionality of our space. It is a method to reduce words to their base forms. We do not apply stemming here, because we want to measure the ideologies' content regarding all vocabulary.

[9] We get this piece of text from a particular speech record (Senator Arthur Virgílio, 53rd legislature, on April 7, 2010.)

[10] Through that, we also avoid words that might be misspelt.

In order to prevent the classifiers from picking up the potential correlations between the names and party affiliations, we also removed senators names from the documents. Our last resource to simplify the data is to identify specific expressions embedded in all documents that not carry significance for our classification method. We notice that certain expressions, such as *Ex$^o$* (Your Excellency) and *srs* (gentlemen), are irrelevant and we remove them. In general, these expressions represent an example of speech standardization.

### *1.3.1 Tf-idf*

A useful approach that excludes both common and rare words is filtering by "term-frequency-inverse-document-frequency", or tf-idf (Gentzkow et al., 2017.) For a word or other feature $j$ in document $d_i$, term frequency ($tf_{ij}$) is the count $w_{ij}$ of occurrences of $j$ in $d_i$. Inverse document frequency ($idf_j$) is the log of one over the share of documents containing $j$: log $(M/s_j)$ where $s_j = \sum_i \mathbb{1}_{\left[w_{ij>0}\right]}$ and $M$ is the total number of documents. The object of interest tf-idf is the product $tf_{ij} \times idf_j$. Thus, very rare words will have low tf-idf scores because $tf_{ij}$ will be low. Moreover, very common words that appear in most or all documents will have low tf-idf scores because $idf_j$ will be low. In other words, tf–idf assigns to term $j$ a weight in document $i$ that is (Manning et al., 2008): i) highest when $j$ occurs many times within a small number of documents (thus lending high discriminating power to those documents); ii) lower when the term occurs fewer times in a document, or occurs in many documents (thus offering a less pronounced relevance signal); and, iii) lowest when the term occurs in virtually all documents.

## 1.4 Supervised Learning

In supervised models, we observe the documents and categories in a training set ($\mathbf{d_{train}}$, $\mathbf{c_{train}}$) such that they may be directly harnessed to inform the model of text generation. During the training, also known as the learning phase, we determine the precise form of the function $f(d_i)$ on the basis of the training data $\mathbf{d_{train}}$. Once we train the model we can then determine the identity of new documents $\mathbf{d_{test}}$ which are said to comprise a test set. The result of running the machine learning algorithm can be expressed as a function $f(d_i)$ which takes a new document $d_i$ as input and that generates an output vector $c_i$, encoded in the same way as the categorical vectors. The ability to categorize correctly new examples that differ from those used for training is known as generalization (Bishop, 2006.)

### *1.4.1 Naive Bayes Classifier*

The most common supervised generative model is the naive Bayes classifier (Gentzkow et al., 2017) which treats counts for each token as independent with class dependent means. In naive Bayes, $c_i$ – our categorical variable – and the token count distribution is factorized as $p(d_i|c_i) = \prod_j p_j(w_{ij}|c_i)$, thus "naively" specifying conditional independence between tokens

$j$ [11]. The parameters of each independent token distribution are estimated, yielding $\hat{p}_j$ for $j = 1, \ldots, p$. We invert the model for prediction, with classification probabilities for the possible class labels obtained via Bayes's rule as

$$p(c|d_i) = \frac{p(d_i|c)\pi_t}{\sum_a p(d_i|a)\pi_a} = \frac{\prod_j p_j(w_{ij}|c)\pi_t}{\sum_a \prod_j p_j(w_{ij}|a)\pi_a}$$

where $\pi_a$ is the prior probability on class $a$ (usually just one over the number of possible classes.) For a full explanation see Appendix A.

### *1.4.2 Validation*

In order to employ a standard supervised learning approach, it is necessary to follow a validation process to ensure that our model is able to make predictions. As our data set is relatively large, we employ a hold-out test validation. This means we withhold some of the sample data from the model identification and estimation process ($\mathbf{d_{train}}$), then use the model to make predictions for the hold-out data $\mathbf{d_{test}}$ in order to see how accurate they are. Following that, we shuffle our speeches prior to dividing them into training and test samples: setting two-thirds of our data set to form the training sample $\mathbf{d_{train}}$ and the remaining one-third to be the test sample $\mathbf{d_{test}}$.

In order to guarantee robustness in our analysis we apply a second validation process: in that case, we observe two measures, Precision and Recall. The former is the fraction of retrieved documents that are relevant, while the latter is the fraction of relevant documents that are retrieved. The measures of precision and recall concentrate the evaluation on the return of true positives, asking what percentage of the relevant documents have been found and how many false positives have also been returned (Manning et al., 2008.)

## 1.5 Dimensionality Reduction

The data set's large number of dimensions – represented by the vocabulary length – limits our understanding when we try to represent their vectors in a high-dimensional space. In order to further comprehend the big picture rendered by our data it is necessary to employ some dimensionality reduction method. Therefore, we apply an algorithm called t-Distributed Stochastic Neighbor Embedding (t-SNE) (van der Maaten and Hinton, 2008) which is an unsupervised, non-linear technique primarily used for data exploration and visualizing high-dimensional data. In simpler terms, t-SNE gives us a feel or intuition of how the data is arranged in a high-dimensional space. That technique is similar to PCA (Principal Component Analysis) which is a linear dimension reduction technique. t-SNE differs from PCA by preserving only

[11] This rules out the possibility that by choosing to say one token (say, "hello") we reduce the probability that we say some other token (say, "hi").

small pairwise distances or local similarities whereas PCA is concerned with preserving large pairwise distances to maximize variance [12].

Before running t-SNE is higly recommended to use another dimensionality reduction method (e.g. PCA for dense data or TruncatedSVD for sparse data) to reduce the number of dimensions to a reasonable amount if the number of features is very high. This will suppress some noise and speed up the computation of pairwise distances between samples (van der Maaten and Hinton, 2008.)[13]

The t-SNE algorithm works calculating a similarity measure between pairs of instances in the high dimensional space ($p_{ij}$) and in the low dimensional space ($q_{ij}$) – for a full explanation see Appendix B.1. It then tries to optimize these two similarity measures using a cost function, which is the Kullback-Leibler divergence (Kullback and Leibler, 1951):

$$C = KL(P||Q) = \sum_i \sum_j p_{ij} log \frac{p_{ij}}{q_{ij}}$$

Finally, we use gradient descent to minimize our KL cost function. The final output will be a two dimensions graph that represents our high-dimensional data. Through its analysis it is possible to perceive the formation of clusters that could illustrate changes in our underlying classification model.

[12] Van der Maaten & Hinton (2008) compare the PCA and t-SNE approaches using the Swiss Roll dataset. They show that due to the non-linearity of this toy dataset (manifold) and preserving large distances that PCA would incorrectly preserve the structure of the data.

[13] In our case, as we are dealing with sparse data we make use of TruncatedSVD in order to reduce the number of our data set's dimensions – then, we set the algorithm (Pedregosa et al., 2011) to reduce our high dimensional data only to 50 dimensions.

## 2 RESULTS AND DISCUSSION

### 2.1 Classification Results

We start our analysis representing each document $d_i$ through their tf-idf values $w_{ij}$ [1]. Then, we proceed by training the classification algorithm through our labeled data set – in that case, our training set $\mathbf{d_{train}}$. Following this classification procedure[2] we are able to test the prediction's accuracy of each training set over its respective test set – concerning the hold-out validation (Tables 1 and 2). In short, we take a document $\mathbf{d_{test \neq train}}$ in our data set and through our learned function $f(w_{ij})$, we are able to generalize about its content. Then, we classify the speech regarding its tf-idf values [3]. This enables us to investigate whether the politicians underlying each ideology employ a similar language pattern during each legislature. We analyze the inter-temporal dynamic of political language in Brazil as well.

Table 1 – Prediction Accuracy on the Training Set Representations

| | $Legislature$ | | | | | |
|---|---|---|---|---|---|---|
| | 50th | 51st | 52nd | 53rd | 54th | 55th |
| $tf - idf$ | 77.1 | 69.1 | 81.1 | 80.9 | 83.2 | 82 |

Table 2 – Hold-out Validation on the Test Set

| | $Legislature$ | | | | | |
|---|---|---|---|---|---|---|
| | 50th | 51st | 52nd | 53rd | 54th | 55th |
| $tf - idf$ | 67.6 | 65.1 | 75.6 | 77 | 79.2 | 74.5 |

The general accuracy throughout the legislatures presents a high average value. Our hold-out process validates it (Table 2.) Although they present values slightly lower than our prediction accuracy on the training set representations, their performance enable us to ensure that our model does not configure overfitting [4]. During the 54th legislature, we find the top average accuracy score among the models. Our results show that most senators employ a truly specific vocabulary, which will be further analyzed in the next sections.

Our data set – as observed through Tables 3 and 4 – presents a very imbalanced distribution among the classes. Looking from the 50th to the 53rd legislatures, we notice that speeches' high frequency is derived from centrist politicians. On the other hand, in our two last legislatures, 54th and 55th, we observe more than 40% of speeches frequency related to leftist senators.

[1] The value of each dimension $j$ is the word frequency normalized by the inverted document frequency, that is, the word frequency divided by the document frequency (the number of documents that contain this word in the whole collection.)

[2] Here, we employ the MultinomialNB algorithm from Pedregosa et al. (2011.)

[3] According to our setup, the document $d_i$ can be classified as $c_i$, s.t. $c_i \in \{-1, 0, 1.\}$

[4] As speech pattern evolves across timeline, we expect that for a model based on previous speeches our classifier tends to present poor performance on the hold-out validation. Therefore, new speeches can bring novelty.

Table 3 – Speech Counting

| | *Legislature* | | | | | |
|---|---|---|---|---|---|---|
| | 50th | 51st | 52nd | 53rd | 54th | 55th |
| Left | 2,784 | 2,498 | 4,833 | 4,932 | 7,594 | 2,605 |
| Center | 4,005 | 5,105 | 9,288 | 7,516 | 6,438 | 2,371 |
| Right | 3,233 | 1,975 | 4,760 | 4,846 | 3,295 | 990 |
| *Total* | 10,022 | 9,578 | 18,881 | 17,294 | 17,327 | 5,966 |

Table 4 – Speech Frequency

| | *Legislature* | | | | | |
|---|---|---|---|---|---|---|
| | 50th | 51st | 52nd | 53rd | 54th | 55th |
| Left | 27% | 26% | 26% | 28% | 44% | 44% |
| Center | 40% | 53% | 49% | 43% | 37% | 40% |
| Right | 33% | 21% | 25% | 29% | 19% | 16% |

Accuracy may not perform well with imbalanced data sets. In that case, we need some method to guarantee the validity of our results. Thus, in order to attribute a robustness check regarding our prediction accuracy analysis, we present the precision-recall curve for each class (and, one average curve) among all legislatures (Figure 1.)[5]

These curves show the trade-off between precision and recall for different thresholds. A high area under the curve represents both high recall and high precision, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate. High scores for both show that the classifier is returning accurate results (high precision), as well as returning a majority of all positive results (high recall.) [6]

Then, we notice a high area under the curve for the 50th, 52nd, 53rd, 54th and 55th legislatures. Although, for the 51st legislature, we observe certain abnormality. Specifically in the 51st legislature we just have a low area for the Right class. Therefore, observing these precision-recall curves, we corroborate with our prediction accuracy analysis. Attesting, then, its validity. It enables us to apply our classification model on the content analysis of each ideology.

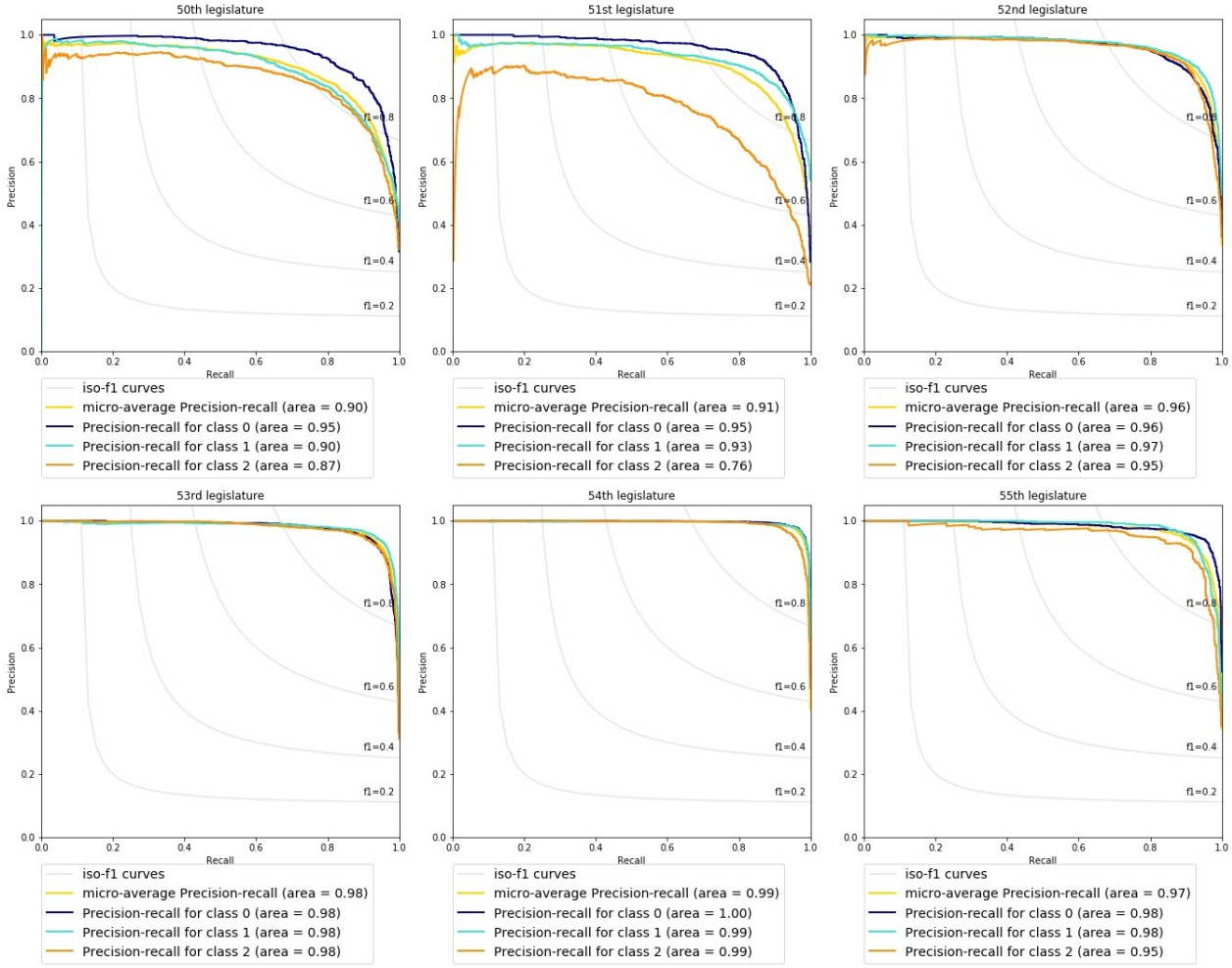
## 2.2 Data Visualization

Analyzing the dynamic across legislatures through our t-SNE visualizations (Figure 2), we notice the emergence of certain discontinuity starting during the 52nd legislature. It suggests that centrist senators after a long period in charge made a political drift. Power & Zucco (2009)

[5] This figure also presents iso-f1 curves, which is defined as the harmonic mean of precision and recall: F1 = $2\frac{P \times R}{P+R}$

[6] A system with high recall but low precision returns many results, but most of its predicted labels are incorrect when compared to the training labels. A system with high precision but low recall is just the opposite, returning very few results, but most of its predicted labels are correct when compared to the training labels.

Figure 1 – Precision-Recall Curves of 50th until the 55th Legislatures Speeches

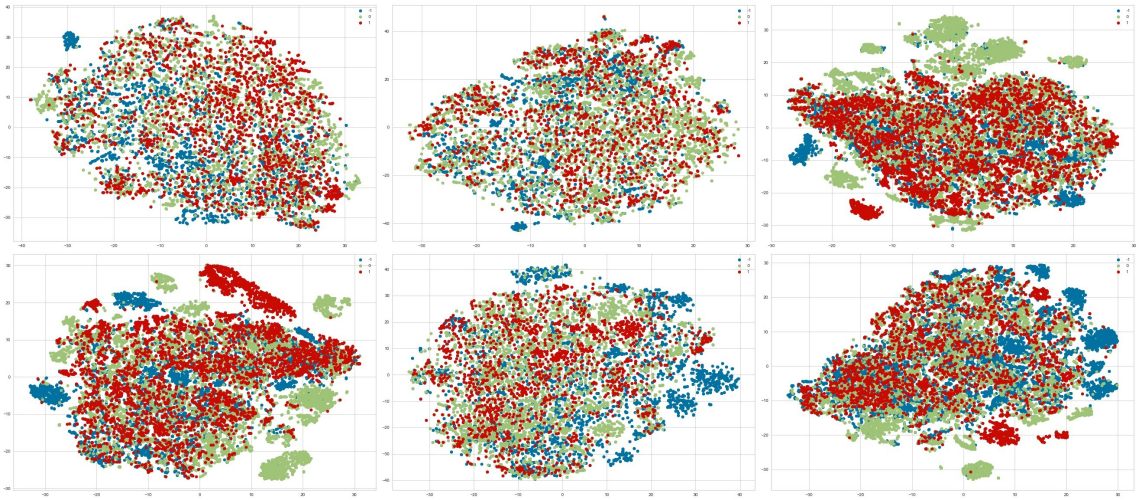



findings corroborate it: the political heritage from Cardoso's government represents this change. According to them, the neoliberal economic reforms implemented by Cardoso's team brought to right-wing politicians certain point of identification. In line with the hypothesis of $direita\ envergonhada$ (literally, the "ashamed right") for a long time most conservative politicians shunned the connection with the right-wing heritage from dictatorship. But now, they have some point of unity.

Analyzing these graphs, we easily perceive that discontinuity. We notice the formation of distinct clusters departing from centrist senators – represented by green color. Despite that new setting, after the 54th legislature we notice that this configuration was not perennial. PSDB's weak bias towards rightism could explains it, in line with Guarnieri (2014.). During the previous decade, PSDB moved itself to a more leftist position (Power & Zucco, 2009.) The adoption of a wider progressivism, measure by their speech patterns, supports that claim – as we show in the next section.

The graphs related to the 54th and the 55th legislatures show the formation of leftist clusters – represented by blue color. We suggest these new clusters stem from discourse convergence

Figure 2 – t-SNE Visualization of 50th to 55th Legislature Speeches



⋆ **Note**: From top-left to top-right: 50th, 51st and 52nd legislatures. From bottom-left to bottom-right: 53rd, 54th and 55th legislatures..

made by left-wing parties. For instance, they joined together to defend Dilma's mandate and the Lula's heritage. Moreover, after Dilma's impeachment, they started a great opposition movement against Temer's government. We further explore it in the next section.

### 2.3 Feature Analysis

Starting from tf-idf representations is possible to visualize dictionaries that best translate each ideology. Through that, we are able to analyze the content of each ideological position. Lets start by visualizing a selection of 15 features of each political spectrum among the top 100 features respectively [7].

Table 5 – Tf-idf Feature Set Analysis for 50th Legislature Vocabulary

| **Left** | **Center** | **Right** |
|---|---|---|
| Terra | Saúde | Desenvolvimento |
| Reforma | Nordeste | Trabalho |
| Pessoas | Municípios | Produção |
| Trabalho | Economia | Agricultura |
| Agrária | Homem | Nordeste |
| Renda | Agrária | Amazônia |
| Cardoso | Terra | Comunicação |
| Respeito | Agricultura | Reforma |

Continued on next page

[7] These features are organized according to their tf-idf weights in a decreasing manner.

**Tabela 5 – continued from previous page**

| Left | Center | Right |
|---|---|---|
| Direitos | Energia | Mercado |
| CPI | Plano | Produtores |
| Crianças | Centro | Agrícola |
| Mulheres | Mercado | Capital |
| Reeleição | Empresas | Rural |
| Desemprego | Dívida | Ensino |
| Previdência | Investimentos | Juros |

Table 6 – Tf-idf Feature Set Analysis for 51st Legislature Vocabulary

| Left | Center | Right |
|---|---|---|
| Renda | Energia | Nordeste |
| CPI | Saúde | Empresas |
| Dívida | Democracia | Energia |
| Salário | Produção | Ensino |
| Internacional | Economia | Agricultura |
| Economia | Agricultura | Produção |
| Dinheiro | Mercado | Água |
| Responsabilidade | CPI | Índios |
| Direitos | Reforma | Universidade |
| Crise | Nordeste | Novo |
| Reforma | Ensino | Conhecimento |
| Pobreza | Violência | Cultura |
| Crianças | Investimentos | Mundial |
| Ética | Tribunal | Reforma |
| Mulheres | Água | Produtores |

In line with our previous section, we notice a subtle difference across political spectrum through the first two legislatures (Tables 5 and 6). The left-wing senators speak more about Rights (in portuguese, *Direitos*), Children (*Crianças*) and Women (*Mulheres*). While centrist and right-wing senators seak more about topics related with production, using words like Debt (*Dívida*), Interest (*Juros*) and Energy (*Energia*). Other interesting feature found through these two periods is the usage of different words to talk about similar topics, showing a distinct perspective towards them. Left-wing senators talk about Agrarian (*Agrária*) and Land (*Terra*),

while right-wing senators employ Agriculture (*Agricultura*). It reveals idiosyncrasies concerning each political ideology, because, in general, rightist parties contemplate politicians derived from country's elites. On the other hand, leftist senators are more concerned with social problematic, such as land reform.

We also notice in the 51st legislature the rightist senators dealing with topics about education, using words such as Knowledge (*Conhecimento*), Culture (*Cultura*) and University (*Universidade*). This represents a paradox, because, in general, the discussion specifically about education is related to progressive politicians. Although, as we see, this topic of discussion was common among conservative senators – revealing, maybe, preoccupation with Brazilian's human capital and its impact in the economic output.

Table 7 – Tf-idf Feature Set Analysis for 52nd Legislature Vocabulary

| **Left** | **Center** | **Right** |
|---|---|---|
| Reforma | Jornal | CPI |
| Saúde | Corrente | Nordeste |
| Crianças | CPI | Deus |
| Mulheres | Saúde | Oposição |
| Previdência | Folha | Segurança |
| CPI | Crescimento | Salário |
| Direitos | Banco | Polícia |
| Crescimento | Dinheiro | Energia |
| Economia | Produção | Tributária |
| PEC | Dirceu | Crescimento |
| Terra | Oposição | Empresas |
| Tributária | Crise | Juros |
| Família | Revista | Economia |
| Polícia | Água | Tribunal |
| Agrária | Nordeste | Corrupção |

Table 8 – Tf-idf Feature Set Analysis for 53rd Legislature Vocabulary

| **Left** | **Center** | **Right** |
|---|---|---|
| Saúde | CPI | Homem |
| Desenvolvimento | Crise | Deus |
| Debate | Tribunal | Vida |
| Direitos | Homem | Nordeste |
| Continued on next page | | |

**Tabela 8 – continued from previous page**

| Left | Center | Right |
|---|---|---|
| Escola | Oposição | CPI |
| Renda | Economia | Justiça |
| Luta | História | Polícia |
| Mulheres | Servidores | Família |
| Aposentados | Dinheiro | Segurança |
| Crianças | Revista | Médico |
| Universidade | CPMF | Crise |
| Salário | Reforma | Pai |
| Professor | Segurança | Parnaíba |
| Violência | Família | Luta |
| Petróleo | Problema | Tribunal |

The presidential victory of left-wing's most important leader, Lula, marks the 52nd legislature (Table 7.) Through this new government, we can notice a substantial change in the vocabulary employed by leftist senators: now, they talk more about conservative topics such as tax and social security reforms (in portuguese, *Tributária* and *Previdência*) [8]. Another difference is the appeal concerning the corruption scheme called *Mensalão* that involved several politicians – many of them from PT, such as former cabinet leader, José Dirceu, and former party president, José Genuíno [9]. Both centrist and right-wing senators have high tf-idf weights for CPI (*Comissão Parlamentar de Inquérito*, or Parliamentary Inquiry Commission) [10], showing the intent to propose an institutional fight against the leftist incumbent government.

The topics regarding land conflicts, revealed through the use of words such as Agrarian and Land, continue to define much of the main characteristics of leftist senators. One interesting point observed through this legislature is the antagonism between left-wing and right-wing senators when they talk about family: as appointed by our tf-idf weights, left-wings senators posit family around the figure of Women and Children, while right-wing senators posit it around the figure of Man (*Homem*). It exposes a view of society for conservative politicians centered on the male figure. We also notice the emergence of high tf-idf weights related to right-wing vocabulary such as God (*Deus*), Safety (*Segurança*) and Police (*Polícia*) – showing us a true conservative pattern.

Specifically this legislature is special in our analysis, because as we show in our previous section it represented an inflection point. Indeed, through our feature analysis we notice

[8] As noted by Power & Zucco (2009): the general trend is that both parties (PT and PSDB) moved markedly to the right while in government and tended to move to the left while in opposition.

[9] Even though others politicians derived from distinct political spectrum were also involved in this bribery scheme, even elite members from right-wing parties such as PTB and PP.

[10] During the 52nd legislature a joint parliamentary inquiry commission was installed involving both the Chamber of Deputies and the Federal Senate in order to investigate the corruption scandals.

a change in the topics discussed. Power & Zucco (2009) pointed that incumbent politicians tend to a more conservative agenda, in our analysis marked by discussions regarding structural reforms. However, the political drift made by PSDB is not easily perceived through this feature analysis. We can just point that the centrist politicians employ a "blurry"vocabulary, sometimes approaching a discussion with more conservative words and sometimes adopting a more progressive jargon. In general, we conclude that their main focus is to talk about the corruption schemes. Thus, the polarization is more explained by electoral results and by political scandals involving the government.

Analyzing the 53rd legislature (Table 8), we notice a certain lack of variability. The trend measured through the feature analysis indicates that topics of discussion in general are the same. Even though we observe subtle differences. The first is the adoption of a more progressive vocabulary by left-wing senators, they started to talk about Education (*Educação*), School (*Escola*), University (*Universidade*), Professor and Wage (*Salário*.) Secondly, the leftist parties members seemed to dodge from discussions involving the corruption scandals, while the others political spectrum revealed a continuity towards those discussions.

Table 9 – Tf-idf Feature Set Analysis for 54th Legislature Vocabulary

| **Left** | **Center** | **Right** |
|---|---|---|
| Mulheres | Minoria | Trabalhista |
| Direitos | Economia | Força |
| Violência | Tribunal | Dinheiro |
| Renda | Corrupção | Médicos |
| Crianças | Poder | Família |
| Nordeste | Supremo | Segurança |
| Liberdade | Documento | Energia |
| Crescimento | Verdade | Indústria |
| Socialista | Crescimento | Produtores |
| Energia | Cidade | Homem |
| Médicos | Empresas | Deus |
| Escola | Produção | Polícia |
| Reforma | Oposição | Mulheres |
| Universidade | Obras | Fronteira |
| Plano | CPI | Infraestrutura |

Table 10 – Tf-idf Feature Set Analysis for 55th Legislature Vocabulary

| **Left** | **Center** | **Right** |
|---|---|---|
| Dilma | Crise | Câncer |
| Lula | Reforma | Mulheres |
| Mulheres | Mulheres | Previdência |
| Impeachment | Dinheiro | Produção |
| Socialismo | Socialismo | Dinheiro |
| Reforma | Energia | Agricultura |
| Golpe | Progressista | Família |
| Direitos | Água | Empresas |
| Violência | Justiça | Polícia |
| Crime | Previdência | Direitos |
| Luta | Impeachment | Trabalhista |
| PEC | Família | Trabalhadores |
| Fiscal | Trabalhadores | Impeachment |
| Previdência | Fiscal | Democrática |
| Michel | Juros | Econômica |

The 54th and the 55th legislatures are truly distinctive – Tables 9 and 10, respectively. While this first legislature was marked by PT continuity in charge – through Dilma Rousseff's victory – with certain economic stability, this second and last legislature was marked by her impeachment process and by a drastic economic crisis. These facts are easily perceived through our feature analysis.

The 54th legislature shows us a continuity with topics previously discussed by senators. While the last legislature already exposed an intense conflict at stake. It is observed through the employ of words such as Coup (*Golpe*), Impeachment and Justice (*Justiça*.) The general motivation of this political battle is observed through usage of words such as Fiscal and Socialism (*Socialismo*), because according to political opinion the principal menaces to democracy were the crimes of responsibility (in general, in the fiscal area) and the slant to "socialist"agenda. One interesting feature noticed here is the proximity between the tf-idf weight values for the words Dilma and Lula. That is, in general, when senators initiated a discussion about Dilma's government they almost always invoked the presence of former president Lula. This pattern is only perceived at this specific case.

After Dilma's impeachment, her vice president assumed the government and tried to implement a reformist agenda. It is observed through usage of words such as PEC (*Proposta de Emenda à Constituição*, or Proposed Amendment to the Constitution) and Social Security (*Previdência*) – this last was widely employed by all the political spectrum. As the Brazilian

economy was impacted by a terrible crisis in the 55th legislature, the necessity to discuss topics related to the economy is revealed by this feature analysis.

In the literature, we find two groups of results. The first, based on low-dimensional representations, indicates the economic dimension as the axis which most distinctly separate political ideologies – as we observe in Poole and Rosenthal (1985.) While the second, based on high-dimensional representations, proposes that moral topics are the dimensions where ideologies are more different – as we observe in Diermeier et al. (2011) and Evans (2003.) According to our text as data analysis, we indeed notice that moral topics define the majority of ideologies.

# 3 CONCLUSION

Through our classification measures, we confirm the hypothesis that ideologies work as belief systems. Indeed, the vocabulary pattern that politicians use in their discussions seemed to represent a great constraint. In general, left-wing senators employ a distinct jargon to discuss topics related to human rights, land reform and education. While right-wing senators have their specific vocabulary to approach topics such as safety and production. These topics do not change a lot across our analyzed interval, although the way in which they are approached seemed to change – with incumbent politicians generally talking about conservative issues.

Regarding the party setup, the inflection point related to polarization is the 52nd legislature with Lula's presidential victory and *Mensalão* scandal. Although, it seemed to be an outcome more tied with electoral results and the coalescence around the issue of corruption than to ideological disputes. Guarnieri (2014), analyzing data from ESEB[1], has found that was a great vacuum on the right of the political spectrum. And, this political vacuum was generally occupied by centrist PSDB, even though this party did not hold the conservative core agenda. This lack of variability through the two next legislatures confirms the hypothesis that our party configuration does not represent the true polarization dynamic revealed in other instances, such as social media[2]. It is also explained by political convergence during the PT governments. Their cabinets were composed by politicians derived from all over the political spectrum, from far-left PC do B to far-right PP. These large government bases made political debate converge to similar points, even though through our data we notice a relative difference among them – specifically in the way that they are approached (their word usage). At the 55th legislature, we already notice a change in the political scenario. With economic crisis and corruption scandals that impeached former president Dilma, left-wing senators seemed to reorganize into a new form to approach the issues in the Parliament.

As the party setting in Brazil does not follow an ideological rationality, it surely represents a weakness in our analysis. Because, as we rely on party membership to assign one ideological label, this might lead to some misrepresentations. Therefore, it proposes a challenge and also it establishes an alternative in order to decode the ideological patterns subsumed in the political speeches.

[1] The Brazilian Electoral Study (ESEB), national post-election survey.

[2] Data from facebook shows a great polarization among Brazilian society: <http://theconversation.com/mapping-brazils-political-polarization-online-96434>.

## REFERENCES

Alston, L. J., M. Melo, B. Mueller and C. Pereira (2013.) Changing Social Contracts: Beliefs and Dissipative Inclusion in Brazil. *Journal of Comparative Economics*, 41(1), 48-65.

Barron, A., J. Huang, R. L. Spang, and S. DeDeo (2018.) Individuals, institutions, and innovation in the debates of the french revolution. *Proceedings of the National Academy of Sciences*, 115(18):4607–4612.

Bishop, C. M. (2006.) *Pattern Recognition and Machine Learning*. Springer.

Bollen, J., H. Mao, and X. Zeng (2011.) Twitter mood predicts the stock market. *Journal of Computational Science* 2(1), 1–8.

Converse, P. E. (1964.) The Nature of Belief Systems in Mass Publics. In *Ideology and Discontent*, edited by D.E. Apter. New York: Free Press.

Diermeier, D., J. F. Godbout, B. Yu and S. Kaufmann (2011.) Language and Ideology in Congress. *British Journal of Politics*, 42, pp. 31-55.

Evans, J. (2003.) Have Americans' Attitudes Become More Polarized? - An Update. Social Science Quarterly. Vol. 84, Nº 01. pp. 71-90

Federal Senate of Brazil (2019.) Speech data concerning the 50th legislature until the 55th legislature (1995-2018):<https://www12.senado.leg.br/dados-abertos/legislativo/plenario/pronunciamentos-de-senadores>.

Gentzkow, M., B. T. Kelly and M. Taddy (2017.) Text as data. NBER Working Paper No. 23276.

Gentzkow, M. and J. M. Shapiro, M. Taddy (2019.) Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech. NBER Working Paper No. 22423.

Graf, R. and R. van der Vossen (2013.) Bits versus brains in content analysis: comparing the advantages and disadvantages of manual and automated methods for content analysis. *Communications : The European Journal of Communication Research*, 38(4): pp. 433-443.

Grimmer, J. and B. M. Stewart (2013.) Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3): pp. 1-31.

Guarnieri, F. (2014.) Comportamento eleitoral e estratégia partidária nas eleições presidenciais no Brasil (2002-2010). Opinião Pública, 2014, vol. 20 (2): 157-177

Harari, Y.N. (2018.) *21 Lessons for the 21st Century*. Spiegel & Grau.

Jensen, J., Naidu, S., Kaplan, E. and L. Wilse-Samson (2012.) Political polarization and the dynamics of political language: Evidence from 130 years of partisan speech.*Brookings Papers on Economic Activity*: 1–81.

Jurafsky, D. and J. H. Martin (2018.) *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Third Edition draft.

Kelly, B., D. Papanikolaou, A. Seru, and M. Taddy (2019.) Measuring technological innovation over the long run. Working Paper.

Iter, D., Yoon, J. and D. Jurafsky (2018.) Automatic Detection of Incoherent Speech for Diagnosing Schizophrenia. *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*. pp. 136-146.

Lakoff, G. (2004.) *Don't think of an elephant! Know your values and frame the debate the essential guide for progressives*. White River Junction, VT: Chelsea Green.

Laver, M., Benoit, K. and J. Garry (2003.) Extracting policy positions from political texts using words as data. *American Political Science Review*, 97(2): 311–331.

Levitsky, S. and D. Ziblatt (2018.) *How democracies die*. New York: Crown.

Kullback, S. and R. Leibler (1951.) On information and sufficiency. *Ann Math*, Stat 22:79–86

Lauderdale, B. E. and A. Herzog (2016.) Measuring political positions from legislative speech. *Political Analysis*, 24(3): pp. 374-394.

Mosteller, F. and D. L. Wallace (1963.) Inference in an authorship problem. *Journal of the American Statistical Association* 58(302), 275–309.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay (2011.) Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, pp. 2825-2830.

Poole, K. T. and H. Rosenthal (1985.) A Spatial Model for Legislative Roll Call Analysis. *American Journal of Political Science*. Vol. 29, Nº 02, pp. 357–384.

Poole, K. T. (2007.) Changing Minds? Not in Congress. *Public Choice*, Vol. 131, pp. 435–51.

Power, T. J. and C. Zucco Jr (2009.) Estimating Ideology of Brazilian Legislative Parties, 1990 - 2005. *Latin American Research Review*, Vol. 44, No. 1.

Sagi, E. and M. Dehghani (2013.) Measuring Moral Rhetoric in Text. *Social Science Computer Review*, Vol. 32 (2), pp. 132-144.

Sagi, E. and D. Diermeier (2015.) Language Use and Coalition Formation in Multiparty Negotiations. *Cognitive Science*, Vol. 41, pp. 259-271.

Scott, S. and H. Varian (2014.) Predicting the present with Bayesian structural time series. *International Journal of Mathematical Modeling and Numerical Optimization* 5(1-2), 4–23.

Tetlock, P. (2007.) Giving content to investor sentiment: the role of media in the stock market. *Journal of Finance* 62(3), 1139–1168

van der Maaten, L. and G. Hinton (2008.) Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605.

Yu, B., S. Kaufmann and D. Diermeier (2008.) Classifying Party Affiliation from Political Speech. *Journal of Information Technology  Politics*, Vol. 5(1).

# APPENDICES

# A NAIVE BAYES CLASSIFIER

## A.1 Multinomial Naive Bayes Classifier

Naive Bayes is a probabilistic classifier, meaning that for a document $d$, out of all classes $c \in C$ the classifier returns the class ĉ which has the maximum posterior probability given the document (Jurafsky and Martin, 2018):

$$\hat{c} = \underset{c \in C}{argmax}\, P(c|d) \quad \text{(A.1)}$$

The intuition of Bayesian classification is to use Bayes' rule to transform Eq. 1 into other probabilities that have some useful properties. Bayes' rule is presented in Eq. 2; it gives us a way to break down any conditional probability $P(x|y)$ into three other probabilities:

$$p(x \mid y) = \frac{p(y \mid x)p(x)}{p(y)} \quad \text{(A.2)}$$

We can then substitute Eq. 2 into Eq. 1 to get Eq. 3:

$$\hat{c} = \underset{c \in C}{argmax}\, P(c \mid d) = \underset{c \in C}{argmax} \frac{P(d \mid c)P(c)}{P(d)} \quad \text{(A.3)}$$

We can conveniently simplify Eq. 3 by dropping the denominator $P(d)$. This is possible because we will be computing $\frac{P(d|c)P(c)}{P(d)}$ for each possible class. But P$(d)$ doesn't change for each class; we are always asking about the most likely class for the same document $d$ which must have the same probability $P(d)$. Thus, we can choose the class that maximizes this simpler formula:

$$\hat{c} = \underset{c \in C}{argmax}\, P(c \mid d) = \underset{c \in C}{argmax} P(d \mid c)P(c) \quad \text{(A.4)}$$

We thus compute the most probable class ĉ given some document $d$ by choosing prior the class which has the highest product of two probabilities: the prior probability of the class $P(c)$ and the likelihood of the document $P(d \mid c)$:

$$\hat{c} = \underset{c \in C}{argmax}\, P(c \mid d) = \underset{c \in C}{argmax} P(d \mid c)P(c) \quad \text{(A.5)}$$

Without loss of generalization, we can represent a document $d$ as a set of features $f_1, f_2, \ldots, f_n$:

$$\hat{c} = \underset{c \in C}{argmax}\, P(f_1, f_2, \ldots, f_n \mid c)P(c) \quad \text{(A.6)}$$

Unfortunately, Eq. 6 is still too hard to compute directly: without some simplifying assumptions, estimating the probability of every possible combination of features (for example, every possible set of words and positions) would require huge numbers of parameters and impossibly large training sets.

The first is the bag of words assumption discussed intuitively above: we assume position doesn't matter, and that the word "love" has the same effect on classification whether it occurs as the 1st, 20th, or last word in the document. Thus we assume that the features $f_1, f_2, \ldots, f_n$ only encode word identity and not position.

The second is commonly called the naive Bayes assumption: this is the conditional independence assumption that the probabilities $P(f_i \mid c)$ are independent given the class $c$ and hence can be 'naively' multiplied as follows:

$$P(f_1, f_2, \ldots, f_n \mid c) = P(f_1 \mid c) \times P(f_2 \mid c) \times \cdots \times P(f_n \mid c) \quad \text{(A.7)}$$

The final equation for the class chosen by a naive Bayes classifier is thus:

$$c_{NB} = \underset{c \in C}{argmax} P(c) \prod_{f \in F} P(f \mid c) \quad \text{(A.8)}$$

Naive Bayes calculations, like calculations for language modeling, are done in log space, to avoid underflow and increase speed. Thus Eq. 9 is generally instead expressed as:

$$c_{NB} = \underset{c \in C}{argmax} logP(c) + \sum_{i \in positions} logP(w_i \mid c) \quad \text{(A.9)}$$

## A.2 Training the Multinomial Naive Bayes Classifier

How can we learn the probabilities $P(c)$ and $P(f_i \mid c)$? For the document prior $P(c)$ we ask what percentage of the documents in our training set are in each class $c$. Let $N_c$ be the number of documents in our training data with class $c$ and $N_doc$ be the total number of documents. Then:

$$\hat{\text{P}}(\text{c}) = \frac{N_c}{N_{doc}} \quad \text{(A.10)}$$

To learn the probability $P(f_i \mid c)$, we'll assume a feature is just the existence of a word in the document's bag of words, and so we'll want $P(w_i \mid c)$, which we compute as the fraction of times the word $w_i$ appears among all words in all documents of topic $c$. We first concatenate all documents with category $c$ into one big "category $c$" text. Then we use the frequency of $w_i$ in this concatenated document to give a maximum likelihood estimate of the probability:

$$\hat{\text{P}}(\text{w}_i \mid c) = \frac{count(w_i, c)}{\sum_{w \in V} count(w, c)} \quad \text{(A.11)}$$

Here the vocabulary $V$ consists of the union of all the word types in all classes, not just the words in one class $c$.

# B DIMENSIONALITY REDUCTION

## B.1 t-SNE

A dataset $\mathcal{X} = x_1, \ldots, x_N$ is a sequence of observations, which are X-dimensional real vectors. The goal of t-SNE is to compute a sequence of points (projection) $\mathcal{Y} = y_1, \ldots, y_N$ where the neighborhoods from $\mathcal{X}$ are preserved, considering that each $y_i \in \mathbb{R}^d$ corresponds to $x_i \in \mathbb{R}^D$. Typically, d = 2 and D $\gg d$.

We start by converting the high-dimensional Euclidean distances between datapoints into conditional probabilities that represent similarities. The similarity of datapoint $x_j$ to datapoint $x_i$ is the conditional probability, $p_{j|i}$, that $x_i$ would pick $x_j$ as its neighbor if neighbors were picked in proportion to their probability density under a Gaussian centered at $x_i$. For nearby datapoints, $p_{j|i}$ is relatively high, whereas for widely separated datapoints, $p_{j|i}$ will be almost infinitesimal (for reasonable values of the variance of the Gaussian, $\sigma_i$.) Mathematically, the conditional probability $p_{j|i}$ is given by (van der Matten and Hinton, 2009):

$$p_{j|i} = \frac{exp(-\|x_i - x_j\|^2/2\sigma_i^2)}{\sum_{k \neq i} exp(-\|x_i - x_k\|^2/2\sigma_i^2)} \quad \text{(B.1)}$$

except for i = j, when $p_{j|i}$ = 0. Each parameter $\sigma_i$ > 0 (where $\sigma_i$ is the variance of the Gaussian that is centered on datapoint $x_i$) is chosen in such a way that the perplexity $\mathcal{K} = 2^{H(P_i)}$ matches a pre-defined value, where $H(P_i)$ is the Shannon entropy of $P_i$ measured in bits: $H(P_i) = -\sum_j p_{j|i} log_2 p_{j|i}$.

Consider also a distinct random process where the probability of choosing a pair $(x_i, x_j) \in \mathcal{X} \times \mathcal{X}$ is defined as:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N} \quad \text{(B.2)}$$

.

Intuitively, $p_{ij}$ is high whenever $p_{j|i}$ or $p_{i|j}$ is high.

In $\mathbb{R}^d$, the probability of choosing a pair $(y_i, y_j) \in \mathcal{Y} \times \mathcal{Y}$ in yet another random process is defined as:

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_k \sum_{l \neq k} (1 + \|y_k - y_l\|^2)^{-1}} \quad \text{(B.3)}$$

except for i = j, when $q_{ij} = 0$. Clearly, $q_{ij}$ is high whenever $y_i$ and $y_j$ are close.

t-SNE aims at minimizing the following cost $C$ with respect to $\mathcal{Y}$:

$$C = KL(P||Q) = \sum_i \sum_j p_{ij} log \frac{p_{ij}}{q_{ij}} \text{ (B.4)}$$

For our purposes, it suffices to note that $C$ corresponds to the Kullback-Leibler divergence between $p_{ij}$ and $q_{ij}$, which heavily penalizes $p_{ij} \gg q_{ij}$, i.e., placing neighbors in $\mathcal{X}$ far apart in $\mathcal{Y}$.

The gradient of $C$ with respect to a point $y_i \in \mathcal{Y}$ is given by:

$$\bigtriangledown_{y_i} C = \frac{\delta C}{\delta y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j)(1 + \|y_i - y_j\|^2)^{-1} \text{ (B.5)}$$

Geometrically, $\bigtriangledown_{y_i} C$ is a combination of vectors pointing in the direction $y_i - y_j$, for every $j$. Each vector $y_i - y_j$ is also weighted by whether $y_j$ should be moved closer to $y_i$ to preserve neighborhoods from $\mathcal{X}$, and by whether $y_j$ is close to $y_i$.

The cost $C$ is usually minimized with respect to $\mathcal{Y}$ by (momentum-based) gradient descent: from an arbitrary initial $\mathcal{Y}$, for a number of iterations, each $y_i \in \mathcal{Y}$ is moved in the direction $- \bigtriangledown_{y_i} C$. For further explanation see van der Maaten & Hinton (2009.)